Using Topic Segmentation Models for the Automatic Organisation of MOOCs Resources

Ghada Alharbi
Department of Computer Science
Sheffield University
Sheffield, S1 4DP, UK
galharbi1@sheffield.ac.uk

Thomas Hain
Department of Computer Science
Sheffield University
Sheffield, S1 4DP, UK
t.hain@sheffield.ac.uk

ABSTRACT

As online courses such as MOOCs become increasingly popular, there has been a dramatic increase for the demand for methods to facilitate this type of organisation. While resources for new courses are often freely available, they are generally not suitably organised into easily manageable units. In this paper, we investigate how state-of-the-art topic segmentation models can be utilised to automatically transform unstructured text into coherent sections, which are suitable for MOOCs content browsing. The suitability of this method with regards to course organisation is confirmed through experiments with a lecture corpus, configured explicitly according to MOOCs settings. Experimental results demonstrate the reliability and scalability of this approach over various academic disciplines. The findings also show that the topic segmentation model which used discourse cues displayed the best results overall.

1. INTRODUCTION

In recent years, Massive Open Online Courses (MOOCs) have been in the spotlight of the media, education professionals, entrepreneurs and technologically aware members of society. As a result, leading universities have been convinced to run their courses online, by establishing open learning platforms, as seen with MIT Open Course Ware (OCW)[1] and Open Yale Courses (OYC)[2].

The majority of these learning platforms organise their resources in line with a pedagogical model, which will allow easy online browsing and accessing [23]. On the other hand, organising these resources takes a great amount of time and is platform dependent, and a large percentage of these platforms have varying formats and structures of the pedagogical model they are based on [23]. In order to decrease the above efforts, unstructured text can be automatically split into coherent sections, which are thus more suitable for online browsing. As these sections include the content of the learning units, an automatic pedagogical annotation model can be employed to organise these units into introductions, descriptions, explanations, examples and other pedagogically significant notions, as examined by [14]. Even though the use of automatic pedagogical annotation models appears suitable, a number of MOOCs sources are structured in line with both the pedagogical and topical approaches. An example of this would be Figure 1(a), the physics lecture from OYC, which displays both ways of structuring. The first and fifth sections depict the pedagogical elements, while the remainder includes the topic segments. This can also be seen in the economics lecture in Figure 1(b).

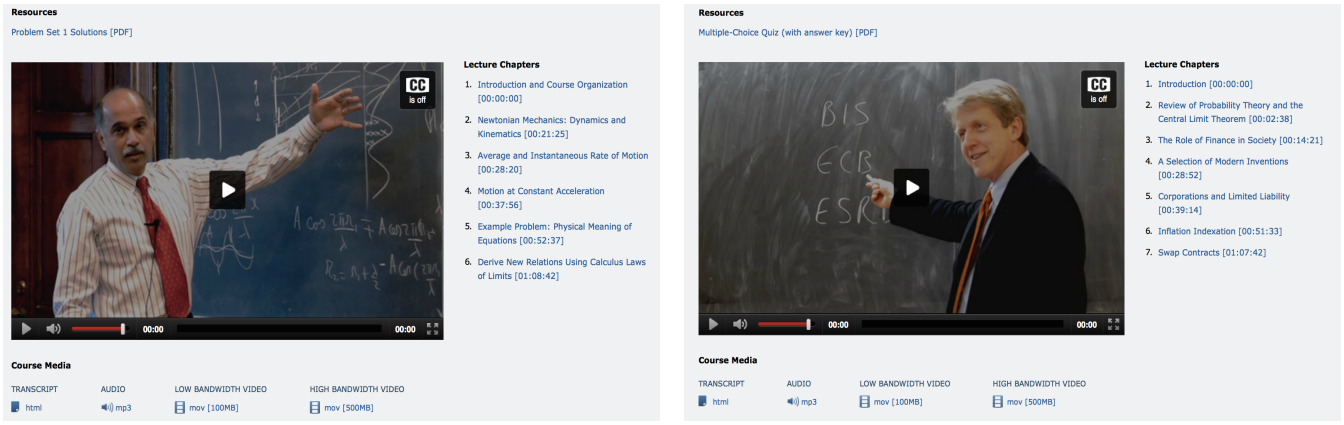
This paper will examine the use of state-of-the-art topic segmentation models to structure lecture resources into cohesive segments, making them suitable for MOOCs content browsing. To evaluate the segmenting applications in the proposed scenario, a test corpus was established using two different disciplines, which were physics and economics, derived from the OYC platform [25, 21]. The topic segmentation models employed in this research include similarity-based models, as seen in [3, 16, 11], language model-based, such as [7, 28] and topic model-based, as seen with [6, 22]. The key strengths of this methodology are its discipline and platform neutrality, which are highlighted in the results of this study. Furthermore, the impact of lexical and discourse cues were examined as features of the segmentation model.

It can be seen from the outcomes that the topic segmentation model which used discourse cues, together with lexical features, showed superior results for the two disciplines. This is due to the fact that discourse cues are often employed to signal the lecturer's aim of the discourse, which means that their learning units are represented more effectively [9]. Despite this, further analysis is required, since the current topic segmentation models hypothesise that discourse cues occur only at the start of an utterance, as seen in [18, 11, 7]. However, other studies have noted that discourse cues can occur at any point in an utterance, and they are a small part of a larger linguistic expression of a writer or speaker [27, 5].

2. BACKGROUND

A number of studies have shown how an automatic pedagogical annotation can be applied to organise lectures resources [14, 24]. However, instead of introducing new aspects such as pedagogical concepts, this paper examined the

[1]http://ocw.mit.edu/index.htm
[2]http://oyc.yale.edu



(a) (b)

Figure 1: Examples of the Interface used for browsing (a) physics and (b) economics lectures in OYC [21, 25].

wider applicability of topic segmentation models for structuring MOOCs content into cohesive units suitable for browsing. In turn, this section concentrates on the work of topic segmentation models, and specifically unsupervised topic segmentation, for either written or spoken language. There has been extensive research on unsupervised segmentation of text, based on lexical cohesion, but certain studies tried to involve other elements, such as discourse or visual cues [7, 8]. This paper will focus mostly on how lexical cohesion is modeled either as similarity-based, language model-based or topic model-based.

TextTiling [12] is considered the first similarity-based model to calculate the cosine similarity between two adjacent blocks of words based purely on word frequency. C99 [3] is based on divisive clustering with a matrix-ranking scheme, while LCSEG modeled lexical chain repetitions of a given lexical term, throughout a fixed-length window of sentences and then chose segmentation points at the local maxima of the cohesion function [11]. MCS [16] optimised normalised minimum-cut criteria, centred on a variation of the cosine similarity between sentences.

An early language model-based algorithm, UI, has been proposed by [28], who tried to find segmentations with compact language models. Furthermore, [7] employs a generative Bayesian model BSEG for topic segmentation. The algorithm computes the maximum likelihood estimates by looking at the entire sequence of sentences, at specific topic boundaries. Also, the model utilises the initial of the potential boundary utterances as discourse cues for the unsupervised model, which is an extension of the work by [11], who automatically identified discourse cues using true labeled boundaries in a supervised fashion.

Latent Dirichlet Allocation (LDA) is a generative model which uses latent structures to model the underlying similarities among observations and it is widely adopted in text analysis to model the shared topics among documents [2]. Topic model-based segmentation was initially interpreted by [26] and built upon by [17]. The most recent LDA based segmenter is TopicTiling [22], which undertakes linear topic segmentation with a pre-trained LDA topic model and estimates the similarity between segments to evaluate text coherence, based on a topic vector representation with cosine similarity. Only the most common topic ID is given to every word in a sentence through Gibbs sampling, in order to maintain efficiency. [6] have shown a hierarchical Bayesian model, which makes use of both Bayesian segmentation and structured topic modelling STM. Superior performance over various models, in both written and spoken texts [6], has been seen with this model. Likewise, the segmentation method of PLDA [20] samples segment boundaries, but also jointly samples a topic model.

The applications of topic segmentation models range from information retrieval to topic tracking [13], summarisation [14] and segmentation of multi-party conversations [11, 20].

3. METHODS

3.1 Data Preparation

Under the Creative-Common license, freely accessible lectures on the OYC website are used as data sources. Expert speakers conducted the lectures, and appear as high quality video and audio data, transcripts, subtitles and lecture segmentation on the course website, as part of MOOCs's initiative. Examples of this segmentation in physics and economics lectures are illustrated in Figure 1. High-level structure distinguishes the lecture as shown in the segmentation. These labelled segments boundaries used as the reference dataset to evaluate the models performance. From these data sources, the two distinct disciplines of physics and economics were selected to establish a new dataset. During the preparation of this study, the total sum of lectures was 47, made up of 24 physics lectures and 23 economics lectures. The average number of annotated segments for the physics lectures was 6, whereas it was 7.1 for the economics lectures. Table 1 shows the new dataset's relevant statistics.

3.2 Segmentation Models

The performance of six competitive models from the literature was compared, with regards to organising MOOCs text content: C99 [4]; UI [28]; LCSEG [11]; MCS [16]; BSEG [7]; STM [6]. All models are explained in Section 2. The publicly available executable given by the authors was employed in all cases, except for LCSEG[3].

[3]This software needs a copyright license from `http://www.cs.columbia.edu/nlp/tools.cgi#LCseg`

	#Lect	#Segments Per Lect	#Total Segments	#Total Words	#Sentences
Physics	24	6	144	260k	18k
Economics	23	7.1	172	212k	15k
Overall	47	6.5	316	472k	33k

Table 1: Lecture Corpus Statistics.

	Physics		Economics	
Text Segmenter	P_k	WD	P_k	WD
C99	0.429	0.433	0.419	0.426
UI	0.426	0.442	0.425	0.435
LCSEG	0.387	0.394	0.356	0.388
MCS	0.439	0.446	0.378	0.383
BSEG	0.364	0.385	0.313	0.334
BSEG+DC	**0.359**	**0.379**	**0.309**	**0.328**
STM	0.372	0.396	0.311	0.330

Table 2: Results of the comparison between segmentation models: WD denotes WindowDiff. Both metrics are penalties, so lower scores indicate better performance.

The paper's specified parameter values [11] were used in the case of LCSEG. MCS needs parameter settings to be tuned on a development set. The corpus of this study does not incorporate development sets, and as a result the tuning was undertaken with the configuration given by the author on the lecture transcript corpus [16]. On the other hand, C99 and UI do not need parameter tuning and can be used without any modification [4, 28]. BSEG also do not need any parameter tuning, but priors are re-estimated, as noted in the paper [7]. The STM model 10 randomly initialised Gibbs chains were used, where every chain ran for $30,000$ iterations, with $25,000$ for burn-in. Following this, 200 samples used the discount parameter $a = 0.2$, and $\lambda_0 = \lambda_1 = 0.1$ and the Dirichlet prior is $\alpha = 0.2$ and $\gamma = 0.01$. In all experiments, the number of segments is assumed to have been given beforehand.

3.3 Evaluation Metrics

All experiments are evaluated with regards to the widely utilised P_k [1] and WindowDiff (WD) [19] metrics. Both metrics run a window throughout a document, and evaluate if the sentences on the edges of the window were suitably segmented with regards to one another. WD is stricter because it needs the number of intervening segments between the two sentences to be exactly the same in both the hypothesised and reference segmentations, whereas P_k only checks if the two sentences are in the same segment. P_k and WD are penalties, so lower values show superior performance. [10] has provided the evaluation source code that was being used.

4. RESULTS AND DISCUSSION

The different performances of the six segmenters using P_k and WD values are shown in table 2. Overall, superior results across the two disciplines were seen in the BSEG model, especially with discourse cues, and the gain and fails of each model across the two disciplines were described. It should be highlighted that these models show better performance using P_k and produce less improvement on the WD metric. This is explained in Section 3.

Notably, the output of the MCS model, which produces segmentation as a graph cut problem, for the physics lectures yields 0.439 P_k, which is worse off compared to more straightforward similarity-based models, such as the C99 and LCSEG. Other models, such as UI, which do not specifically depend on pairwise similarity analysis, have better performance ($P_k = 0.426$) in physics lectures, when compared to MCS. UI calculates a better segmentation performance by estimating alterations to the language model predictions through various partitions, as described in Section 2. On the other hand, economics lectures differed, as MCS had superior performance ($P_k = 0.378$) compared to both C99 and UI, which yielded $P_k = 0.419$ and $P_k = 0.425$ respectively. This is due to the difference in distributional properties of the physics lectures, which were not coherent in their thematic shifts and thus caused a level of distributional differences.

A further note from Table 2 with regards to the LCSEG model was that it had superior performance on P_k metric for both disciplines ($P_k = 0.387$ in physics and $P_k = 0.356$ in economics), compared to all other models used with the exception of BSEG and STM. STM achieved favourable performance, especially in economics lectures, and attained results close to the BSEG model in physics lectures. This can be attributed once again to the lack of coherence in physics lectures, which results in smooth distributional variations. A substantial and consistent increase is seen through the use of BSEG+DC for all lecture subjects. This can be justified from the existence of discourse cues, as depicted in the results of $P_k = 0.359$ in physics and $P_k = 0.309$ in economics. As spoken language is more impulsive and not as planned as written language, the speaker must inform the listener of any alterations to topic content, through the introduction of subtle cues, and references to prior topics during topical transitions [9].

A further analysis study of discourse cues was undertaken, using the labelled topic boundaries. For every word in the lecture corpus, the number of its occurrences near any topic boundary (with a window size of 5 seconds on either side of the target boundary, inclusive) are counted, and set against those further away. The findings were utilised in the undertaking of the χ^2 significance test. The chi-square test allows the calculation of the significance of the near-against distinct-statistics by comparing with the overall statistics, where the null hypothesis is assumed. The word with an χ^2 value in opposition to the hypothesis under 0.01-level confidence (the rejection criterion is $\chi^2 \geq 6.635$) were chosen. Table 4 shows discourse cues sorted by chi-squared value, where bold denotes the common cues of both disciplines. The corpus was manually examined to find these automatically selected discourse cues, and it was discovered that these cues establish linguistic expressions, as in the study by [27] on summarisation task. An example of this is the cue "*topic*", which is part of one expression, such as "*The topic of this lecture is*" or a very different expression, like "*Let's move to another topic*". These expressions can obviously show the function and the purpose of the discourse, and thus show the pedagogical element of this segment. However, current topic

	Near	Distant	Near	Distant
today	14	53	29	92
Other	15254	244776	17838	193820

Table 3: $\chi^2 = 24.73$ in Physics and 35.82 in Economics.

Physics		Economics	
DC	χ^2	DC	χ^2
topic	110.25	**talk**	140.07
last	105.78	about	126.15
okay	51.89	wanted	97.61
now	37.88	**lecture**	67.97
next	32.05	let	66.97
today	24.73	**we**	65.46
alright	22.52	move	59.17
lecure	18.84	**today**	35.83
we	11.40	conclude	29.07
talk	8.68	start	21.58

Table 4: Automatically selected discourse cues (DC), sorted by chi-squared χ^2 value at the level of $p < 0.01$. Boldface indicates that these cues are common across disciplines.

segmentation models do not account for these expressions, possibly because of the fact that these models lack conversational analysis. Additional research is required to examine this aspect, including the induction of these expressions in the segmentation model and the possibility of using an automatic method to identify and extract these expressions, such as in the study by [15] on the extraction of expressions from student essays.

5. CONCLUSION AND FUTURE WORKS

The application of topic segmentation models for the automatic organisation of MOOCs resources has been presented above. The manual analysis of these resources shows that their structure is centred on both pedagogical and topical aspects, and so a new corpus has been established based on this scenario, through two different domains. The study employs the different features of the topic segmentation models in order to compare the results. The outcomes show that the topic segmentation model which utilised linguistic cues (e.g. *today*, *okay*) had the highest results. An important element for future research is the automatic detection and extraction of linguistic expressions, which are used to show various purposes and functions in discourse, in order to be able to involve them in the topic segmentation model. It can be hypothesised that this type of model would have superior performance in the representation of MOOCs learning units.

6. REFERENCES

[1] D. Beeferman, A. Berger, and J. Lafferty. Statistical models for text segmentation. *Mach. Learn.*, 34(1-3):177–210, Feb. 1999.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[3] F. Y. Y. Choi. Advances in domain independent linear text segmentation. In *Proceedings of NAACL'00*, pages 26–33, 2000.

[4] F. Y. Y. Choi, P. Wiemer-Hastings, and J. Moore. Latent semantic analysis for text segmentation. In *In Proceedings of EMNLP'01*, pages 109–117, 2001.

[5] R. Correia, N. Mamede, J. Baptista, and M. Eskenazi. Using the crowd to annotate metadiscursive acts. *In Proceedings 10th Joint ISO-ACL SIGSEM*, page 102, 2014.

[6] L. Du, W. L. Buntine, and M. Johnson. Topic segmentation with a structured topic model. In *HLT-NAACL*, pages 190–200, 2013.

[7] J. Eisenstein and R. Barzilay. Bayesian unsupervised topic segmentation. *Proceedings of EMNLP'08*, page 334, 2008.

[8] J. Eisenstein, R. Barzilay, and R. Davis. Gestural cohesion for topic segmentation. In *Proceedings of ACL-08: HLT*, pages 852–860, 2008.

[9] J. Flowerdew and L. Miller. The teaching of academic listening comprehension and the question of authenticity, 1997.

[10] C. Fournier. Evaluating Text Segmentation using Boundary Edit Distance. In *Proceedings of ACL'13*, 2013.

[11] M. Galley, K. McKeown, E. Fosler-Lussier, and H. Jing. Discourse segmentation of multi-party conversation. In *Proceedings of ACL'03*, pages 562–569, 2003.

[12] M. A. Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*, 23(1):33–64, Mar. 1997.

[13] X. Huang, F. Peng, D. Schuurmans, N. Cercone, and S. E. Robertson. Applying machine learning to text segmentation for information retrieval. *Information Retrieval*, 6(3-4):333–362, 2003.

[14] N. Kokhlikyan, A. Waibel, Y. Zhang, and J. Y. Zhang. Measuring the structural importance through rhetorical structure index. In *HLT-NAACL*, 2013.

[15] N. Madnani, M. Heilman, J. Tetreault, and M. Chodorow. Identifying high-level organizational elements in argumentative discourse. In *Proceedings of NAACL'12: HLT*, pages 20–28, 2012.

[16] I. Malioutov and R. Barzilay. Minimum cut model for spoken lecture segmentation. In *Proceedings ACL '06*, pages 25–32, 2006.

[17] H. Misra, F. Yvon, J. M. Jose, and O. Cappe. Text segmentation via topic modeling: An analytical study. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pages 1553–1556, 2009.

[18] R. J. Passonneau and D. J. Litman. Discourse segmentation by human and automated means. *Comput. Linguist.*, 23(1):103–139, Mar. 1997.

[19] L. Pevzner and M. A. Hearst. A critique and improvement of an evaluation metric for text segmentation. *Comput. Linguist.*, 28(1):19–36, Mar. 2002.

[20] M. Purver, T. L. Griffiths, K. P. Körding, and J. B. Tenenbaum. Unsupervised topic modelling for multi-party spoken discourse. In *Proceedings of COLING-ACL'06*, pages 17–24, 2006.

[21] S. Ramamurti. Fundamental of physics 1 (yale university: Open yale courses). `http://oyc.yale.edu/physics/phys-200#overview`, 2006. (Accessed December 20, 2014), License: Creative Commons BY-NC-SA.

[22] M. Riedl and C. Biemann. Topictiling: A text segmentation algorithm based on lda. In *Proceedings of ACL'12 Student Research Workshop*, pages 37–42, 2012.

[23] O. Rodriguez. The concept of openness behind c and x-moocs (massive open online courses). *Open Praxis*, 5(1):67–73, 2013.

[24] K. Sathiyamurthy and T. V. Geetha. Automatic organization and generation of presentation slides for e-learning. *Int. J. Distance Educ. Technol.*, 10(3):35–52, July 2012.

[25] R. J. Shiller. Financial markets (yale university: Open yale courses). `http://oyc.yale.edu/economics/econ-252-11`, 2011. (Accessed December 22, 2014), License: Creative Commons BY-NC-SA.

[26] Q. Sun, R. Li, D. Luo, and X. Wu. Text segmentation with lda-based fisher kernel. In *Proceedings of ACL'08: Short Papers*, pages 269–272, 2008.

[27] S. Teufel and M. Moens. Summarizing scientific articles - experiments with relevance and rhetorical status. *Computational Linguistics*, 28:2002, 2002.

[28] M. Utiyama and H. Isahara. A statistical model for domain-independent text segmentation. In *Proceedings of ACL'01*, pages 499–506, 2001.